Overview of the Albayzin 2010 Language Recognition Evaluation: database design, evaluation plan and preliminary analysis of results

Luis Javier Rodríguez-Fuentes, Mikel Penagarikano, Amparo Varona,
Mireia Díez, Germán Bordel

Software Technologies Working Group (http://gtts.ehu.es)
Department of Electricity and Electronics, University of the Basque Country
Barrio Sarriena s/n, 48940 Leioa, Spain
email: luisjavier.rodriguez@ehu.es

FALA 2010, Vigo, Spain
November 10, 2010

GTTS
Tecnologías Software

Outline

1. Motivation
2. The language detection task
3. Test conditions
4. Data
5. Organization
6. Results
    - Participation
    - CR-30 (mandatory condition)
    - Dependence on duration
    - Open-set tests
    - Performance on noisy speech
7. Post-eval activity
8. Conclusions

Motivation

- To promote collaboration between research groups (specially from Spain and Portugal) interested in automatic language recognition
- To produce speech resources specifically designed for language recognition applications featuring Iberian languages as target languages
- To explore the limits of state-of-the-art technology (and eventually to foster research progress and technological developments) on wide-band speech from TV broadcasts, which are not used in NIST evaluations
- To evaluate performance degradation when dealing with noisy signals

The language detection task

- As for NIST LRE: *given a segment of speech and a language of interest (target language), determine whether or not that language is spoken in the segment, based on an automated analysis of the data contained in the segment.*
- **Trial:** audio segment + target language + set of non-target languages
- **System output:** hard decision + score (maybe LLR)

Test conditions

- **Set of trials**
  - Closed-set tests (C): only trials corresponding to audio segments containing target languages
  - Open-set tests (O): all the trials
- **Background conditions**
  - Clean speech (C)
  - Noisy/Overlapped speech (N)
- **Nominal duration of audio segments:** 30, 10 and 3 seconds
- **Performance measures** (as defined in NIST LRE, using NIST software, see paper for details):
  - C_{avg} ($P_{target} = 0.5$, $C_{miss} = C_{fa} = 1$)
  - C_{LLR}
  - DET curves

Database features (1)

- KALAKA-2 (includes KALAKA in train and development)
- 6 target languages: Basque, Catalan, English, Galician, Portuguese and Spanish
- Other languages (to allow open-set tests): Arabic, French, German and Romanian
- Audio files: 16 kHz, single channel, 16 bits/sample, PCM (WAV)
- Speech signals extracted from TV broadcast recordings, featuring various dialects, linguistic competence levels, speech modalities and diverse environment conditions
- Disjoint subsets of TV shows posted to train, development and evaluation, as an attempt to guarantee speaker independence
- Size: around 125 hours (distributed in 5 DVD)
  - Train dataset > 82 hours (more than 12 hours per target language)
  - Development dataset > 21 hours
  - Evaluation dataset > 21 hours

Database features (2)

- Segments for training had no length restrictions: clean (more than 10 hours per target language) and noisy segments (around 2 hours per target language) were provided
- Segments for development and evaluation:
  - enclosed by a certain amount of low-energy frames
  - 3-second subset ⊂ 10-second subset ⊂ 30-second subset
  - length tolerance: 3-5, 10-12 and 30-33 seconds (30-35 for noisy segments)
- Size of the development and evaluation datasets:
  - Development: 4950 segments (1458 noisy, 1374 OOS)
  - Evaluation: 4992 segments (1647 noisy, 1320 OOS)

Evaluation rules (in brief)

- 4 test conditions (CC, CN, OC, ON) × 3 durations: 12 tracks
- For each test condition: single primary + any number of contrastive systems
- Results in NIST LRE format (text file with one line per trial and 6 fields per line)
- Participants committed to specify whether or not their scores may be interpreted as log-likelihood ratios
- Participants committed to send descriptions of their systems and present them at the Albayzin 2010 LRE Workshop (after this session)
- Systems ranked in each track according to C_{avg}
- **Award:** system yielding the least C_{avg} in the CC-30 condition

Schedule (as finally executed)

🟪 Evaluation plan released, registration opens (deadline: July 15)

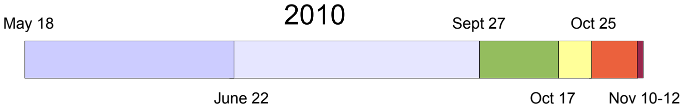
⬜ Train and development data (4 DVD) submitted to registered sites, time for system development

🟩 Evaluation data released, time for processing evaluation data

🟨 System results and descriptions submitted to organization, analysis of the submitted results

🟧 Keyfile and results released, time for preparing final descriptions (deadline: November 2) and workshop presentations

🟥 Albayzin 2010 LRE Workshop (delivery of the 5th DVD: evaluation data and documentation)



Database production

- April-September 2008 (KALAKA, reused for KALAKA-2)
- October-November 2008 + April-May 2010 (train and dev data for new languages)
- August-September 2010 (additional evaluation data)

Participation

Participation: 4 teams, 21 systems

- GTC-VIVOLAB (4 systems: CC, OC: primary, contrastive)
- L^2F (12 systems: all conditions: primary, contrastive-1, contrastive-2)
- UEF-NTNU (1 system: CC: primary)
- UVIGO-GTM (4 systems: CC, CN: primary, contrastive)

Participation

Participation: 4 teams, 21 systems

- GTC-VIVOLAB (4 systems: CC, OC: primary, contrastive)
- L^2F (12 systems: all conditions: primary, contrastive-1, contrastive-2)
- UEF-NTNU (1 system: CC: primary)
- UVIGO-GTM (4 systems: CC, CN: primary, contrastive)

Processing time: all systems below 1×RT

Systems	**CPU-RAM**	**×RT**
GTC-VIVOLAB	–	0.9
L2F	2xQuad Xeon E5530 2.4GHz, 48 GB	0.51
UEF_NTNU	Xeon X5450 3.0GHz	0.051
GTM (p)	Xeon E5620 2.4 GHz, 18 GB	0.0288
GTM (c)	Xeon E5620 2.4 GHz, 18 GB	0.0533

CC-30 (mandatory condition)

C_{avg} for systems submitted to the **CC-30** test condition (in parentheses, results for post-key submissions)

	CC-30		
	primary	**contrastive-1**	**contrastive-2**
GTC-VIVOLAB	**0.0184**	0.0238	–
L^2F	0.0320 (0.0223)	0.0910 (0.0219)	**0.0181**
UEF-NTNU	0.1636	–	–
UVIGO-GTM	0.1916	0.2888	–

CC-30 (mandatory condition)

C_{avg} for systems submitted to the **CC-30** test condition (in parentheses, results for post-key submissions)

	CC-30		
	primary	**contrastive-1**	**contrastive-2**
GTC-VIVOLAB	**0.0184**	0.0238	–
L^2F	0.0320 (0.0223)	0.0910 (0.0219)	**0.0181**
UEF-NTNU	0.1636	–	–
UVIGO-GTM	0.1916	0.2888	–

- **Award winner:** GTC-VIVOLAB (best primary system in CC-30)

CC-30 (mandatory condition)

C_{avg} for systems submitted to the **CC-30** test condition (in parentheses, results for post-key submissions)

	CC-30		
	primary	**contrastive-1**	**contrastive-2**
GTC-VIVOLAB	**0.0184**	0.0238	–
L^2F	0.0320 (0.0223)	0.0910 (0.0219)	**0.0181**
UEF-NTNU	0.1636	–	–
UVIGO-GTM	0.1916	0.2888	–

- **Award winner:** GTC-VIVOLAB (best primary system in CC-30)
- Best result in CC-30: $C_{avg} = 0.0181$ (L^2F contrastive-2)

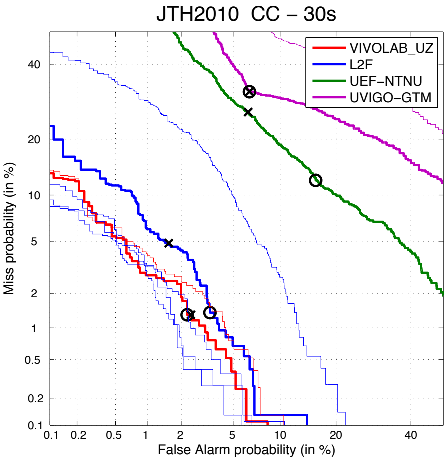
CC-30 (mandatory condition)

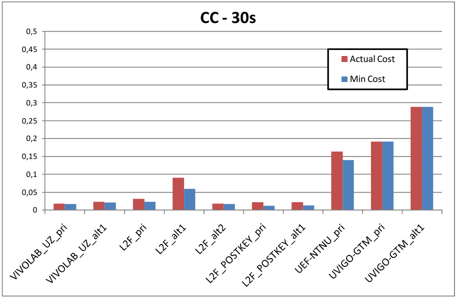
C_{avg} for systems submitted to the **CC-30** test condition (in parentheses, results for post-key submissions)

	CC-30		
	primary	**contrastive-1**	**contrastive-2**
GTC-VIVOLAB	**0.0184**	0.0238	–
L^2F	0.0320 (0.0223)	0.0910 (0.0219)	**0.0181**
UEF-NTNU	0.1636	–	–
UVIGO-GTM	0.1916	0.2888	–

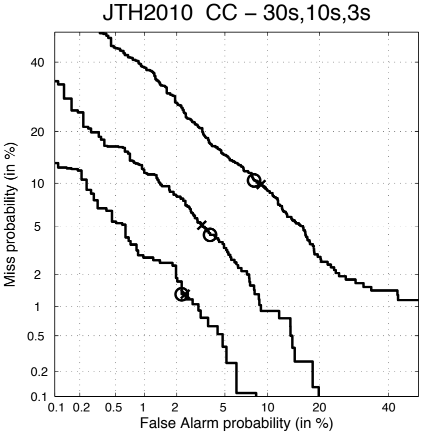
- **Award winner:** GTC-VIVOLAB (best primary system in CC-30)
- Best result in CC-30: $C_{avg} = 0.0181$ (L^2F contrastive-2)
- Post-key submissions from L^2F didn't outperform the two systems above

CC-30 (mandatory condition)



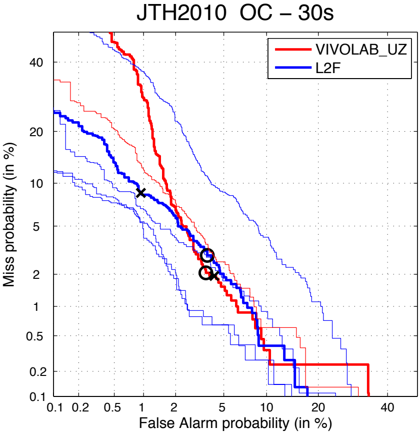


Dependence on duration



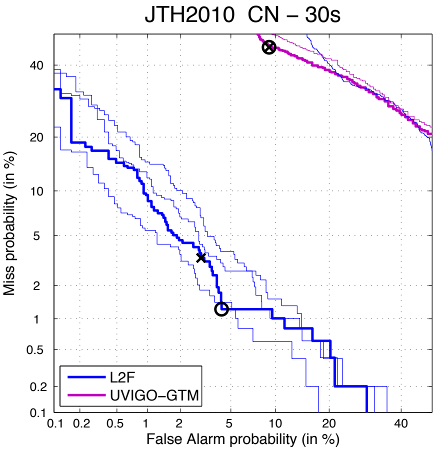
- C_{avg} doubled from 30 to 10, and from 10 to 3 seconds (best primary system in CC-30)
- Similar trend in other conditions and for other systems
- Consistent with previous results in other evaluations

Open-set tests



- $C_{avg} = 0.0307$ for GTC-VIVOLAB(p) in OC-30 (67% cost increase wrt CC-30)
- Similar figures for other systems: 49% and 88% cost increases for L^2F(p) and L^2F(c2)
- Best performance in OC-30: $C_{avg} = 0.0296$ (L^2F primary-postkey)
- As shown in DET curves, C_{min} for some L^2F systems was below 0.02: over-training on dev? bad calibration?

Performance on noisy speech



- New condition in this evaluation: noisy speech
- Only L^2F and UVIGO-GTM submitted systems to this condition
- Surprisingly good performance: cost increases *only* between 30% and 50% wrt performance on clean speech
- L^2F(p) yielded lower cost for CN-30 than for CC-30 !!
- Best performance in CN-30: $C_{avg} = 0.0253$ (L^2F contrastive-2)

Performance on noisy speech

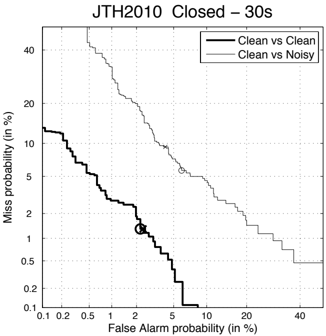
- How do systems designed for clean speech behave when dealing with noisy speech?

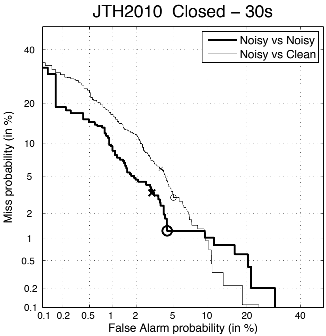
Performance on noisy speech

- How do systems designed for clean speech behave when dealing with noisy speech?
- How do systems designed for noisy speech behave when dealing with clean speech?

Performance on noisy speech

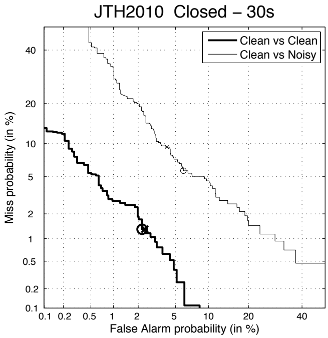
- How do systems designed for clean speech behave when dealing with noisy speech?
- How do systems designed for noisy speech behave when dealing with clean speech?

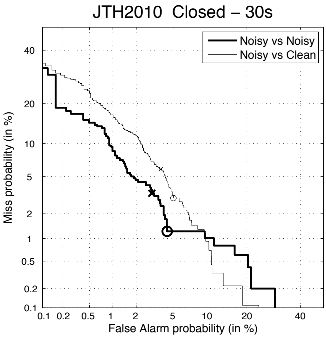




Performance on noisy speech

- How do systems designed for clean speech behave when dealing with noisy speech?
- How do systems designed for noisy speech behave when dealing with clean speech?

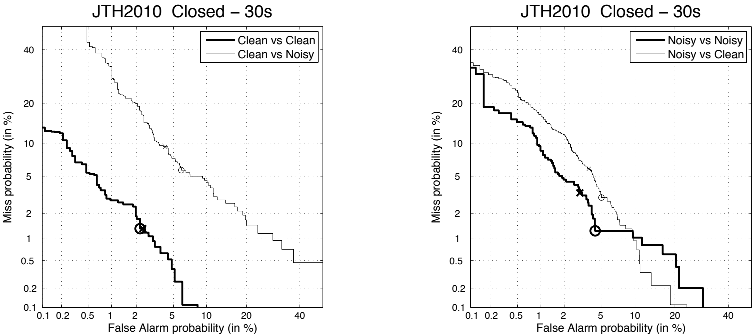




- In the first case (GTC-VIVOLAB): from $C_{avg} = 0.0184$ to $C_{avg} > 0.05$

Performance on noisy speech

- How do systems designed for clean speech behave when dealing with noisy speech?
- How do systems designed for noisy speech behave when dealing with clean speech?



- In the first case (GTC-VIVOLAB): from $C_{avg} = 0.0184$ to $C_{avg} > 0.05$
- In the second (L^2F): from $C_{avg} = 0.0316$ to $C_{avg} \approx 0.05$

Exploring cross-site fusions

- **Proposal:** to investigate which subsystems produce the best combinations under a FoCal-based fusion paradigm

Exploring cross-site fusions

- **Proposal:** to investigate which subsystems produce the best combinations under a FoCal-based fusion paradigm
- Unexplored cross-site fusions may give valuable cues about which kind of systems would be worth developing and combining

Exploring cross-site fusions

- **Proposal:** to investigate which subsystems produce the best combinations under a FoCal-based fusion paradigm
- Unexplored cross-site fusions may give valuable cues about which kind of systems would be worth developing and combining
- Focus on the core condition (CC-30)

Exploring cross-site fusions

- **Proposal:** to investigate which subsystems produce the best combinations under a FoCal-based fusion paradigm
- Unexplored cross-site fusions may give valuable cues about which kind of systems would be worth developing and combining
- Focus on the core condition (CC-30)
- 3 sites submitted log-likelihoods for their subsystems (previously undisclosed)

Exploring cross-site fusions

- **Proposal:** to investigate which subsystems produce the best combinations under a FoCal-based fusion paradigm
- Unexplored cross-site fusions may give valuable cues about which kind of systems would be worth developing and combining
- Focus on the core condition (CC-30)
- 3 sites submitted log-likelihoods for their subsystems (previously undisclosed)
- The organizing team (GTTS) provided the log-likelihoods for 3 subsystems developed for this evaluation

Exploring cross-site fusions

- **Proposal:** to investigate which subsystems produce the best combinations under a FoCal-based fusion paradigm
- Unexplored cross-site fusions may give valuable cues about which kind of systems would be worth developing and combining
- Focus on the core condition (CC-30)
- 3 sites submitted log-likelihoods for their subsystems (previously undisclosed)
- The organizing team (GTTS) provided the log-likelihoods for 3 subsystems developed for this evaluation
- All the information (log-likelihoods and brief descriptions of subsystems) was uploaded and results were released through the wiki

Exploring cross-site fusions

Best cross-site fusions (for n subsystems, $n \in [1,5]$)

n	$C_{LLR}^{(dev)}$	$C_{LLR}^{(eval)}$	$C_{avg}^{(eval)}$	**Best fusion**
1	0.23853	0.20643	**0.0207**	GTTS_CZ
2	0.02662	0.12151	**0.0094**	L2F_PPRLM-ES+UZ_jfa
3	0.02066	0.10831	**0.0066**	L2F_PPRLM-EN+L2F_PPRLM-ES+UZ_jfa
4	0.02707	0.11011	**0.0059**	GTTS_CZ+L2F_PPRLM-ES+UZ_mmi+UZ_PRLM_ru
5	0.01430	0.09723	**0.0054**	GTTS_HU+L2F_PPRLM-ES+UZ_jfa+UZ_ml+UZ_PRLM_hu

The best fusion of 5 subsystems yielded $C_{avg} = 0.0054$, 3 times lower than that obtained by the best system in CC-30 (meaning 70% cost decrease)

Conclusions

ALBAYZIN 2010 Language Recognition Evaluation

- 6 target languages, including all the official languages in Spain and Portugal

Conclusions

$\boxed{\text{ALBAYZIN 2010 Language Recognition Evaluation}}$

- 6 target languages, including all the official languages in Spain and Portugal
- A new database: **KALAKA-2** (125 hours), consisting of 16kHz speech signals taken from TV broadcasts

Conclusions

$\boxed{\text{ALBAYZIN 2010 Language Recognition Evaluation}}$

- 6 target languages, including all the official languages in Spain and Portugal
- A new database: **KALAKA-2** (125 hours), consisting of 16kHz speech signals taken from TV broadcasts
- Best system in the core condition: $\boxed{C_{avg} = 0.0181}$, a remarkable technology improvement with regard to the Albayzin 2008 LRE ($C_{avg} = 0.0552$)

Conclusions

ALBAYZIN 2010 Language Recognition Evaluation

- 6 target languages, including all the official languages in Spain and Portugal
- A new database: **KALAKA-2** (125 hours), consisting of 16kHz speech signals taken from TV broadcasts
- Best system in the core condition: $\boxed{C_{avg} = 0.0181}$, a remarkable technology improvement with regard to the Albayzin 2008 LRE ($C_{avg} = 0.0552$)
- New test condition on **noisy speech**: reasonably good results can be attained if suitable data are available to train and calibrate systems

Conclusions

$\boxed{\text{ALBAYZIN 2010 Language Recognition Evaluation}}$

- 6 target languages, including all the official languages in Spain and Portugal
- A new database: **KALAKA-2** (125 hours), consisting of 16kHz speech signals taken from TV broadcasts
- Best system in the core condition: $\boxed{C_{avg} = 0.0181}$, a remarkable technology improvement with regard to the Albayzin 2008 LRE ($C_{avg} = 0.0552$)
- New test condition on **noisy speech**: reasonably good results can be attained if suitable data are available to train and calibrate systems
- *Post-eval activity*: **cross site *FoCal*-based subsystem fusions** revealed great performance improvements, e.g. best fusion of 5 subsystems yielded $\boxed{C_{avg} = 0.0054}$

GTTS Tecnologías Software

Acknowledgements

- Thanks to the Organizing Committee of FALA 2010, specially Laura Docio for her support

Acknowledgements

- Thanks to the Organizing Committee of FALA 2010, specially Laura Docio for her support
- Thanks to all the participants, for their work and feedback, and their collaboration in the cross-site fusion activity

Acknowledgements

- Thanks to the Organizing Committee of FALA 2010, specially Laura Docio for her support
- Thanks to all the participants, for their work and feedback, and their collaboration in the cross-site fusion activity
- Thank you all for your patience !!